Preface

During the last decade companies, governments, and research groups worldwide have directed significant effort towards the creation of sophisticated digital libraries across a variety of disciplines. As digital libraries proliferate, in a variety of media, and from a variety of sources, problems of resource selection and data fusion become major obstacles. Traditional search engines, even very large systems such as Google, are unable to provide access to the "Hidden Web" of information that is only available via digital library search interfaces. Effective, reliable information retrieval also requires the ability to pose multimedia queries across many digital libraries. The answer to a query about the lyrics to a folk song might be text or an audio recording, but few systems today could deliver both data types in response to a single, simple query. Distributed information retrieval addresses issues that arise when people have routine access to thousands of multimedia digital libraries.

The SIGIR 2003 Workshop on Distributed Information Retrieval was held on August 1, 2003, at the University of Toronto, following the SIGIR conference, to provide a venue for the presentation and discussion of recent research on the design and implementation of methods and tools for resource discovery, resource description, resource selection, data fusion, and user interaction. About 25 people attended, including representatives from university and industrial research labs. Participants were encouraged to ask questions during and after presentations, which they did. The formal presentations were followed by a general discussion of the state-of-the-art in distributed information retrieval, with a particular emphasis on what still needs to be done.

This volume includes extended and revised versions of the papers presented in the SIGIR 2003 Workshop in addition to a few invited papers.

Overview

The book is structured into four parts covering the major areas of research associated with multimedia distributed information retrieval: resource discovery, resource selection, data fusion, and architectures. Below we provide very brief descriptions of the papers included in the volume, to give a sense of the range of themes and topics covered.

Harvesting: broadening the field of distributed information retrieval by Fox et al. argues that in addition to federated search and gathering, harvesting is an important approach to address the needs of distributed information retrieval. The paper is centered on the user of the Open Archives Initiative (OAI) Protocol for Metadata Harvesting. It illustrates the use of this protocol in three projects: OAD, NDLTD, and CITIDEL. The OAI protocol extends traditional digital libraries services in a user-centered fashion. This is exemplified in the ESSEX

search engine, which also enables multischeming browsing and quality-oriented filtering.

Using query probing to identify query language features on the Web by Bergholz and Chidlovskii addresses the problem of discovering the lexical processing and query language characteristics of an uncooperative search engine. Their research shows that a relatively small number of probe queries and simple classification algorithms are sufficient to discover a range of search engine characteristics, including stopword removal, stemming, phrase processing, and treatment of AND operators, much of the time. In evaluations with 19 search engines, features are discovered correctly about 75–80% of the time. Future research will be directed at improving classification accuracy, for example with better feature selection and improved classification algorithms.

The effect of database size distribution on resource selection algorithms by Si and Callan extends work reported in the main SIGIR conference. The conference paper reported on a new resource selection algorithm (ReDDE) that compensates for skewed distributions of database sizes more effectively than prior algorithms. The workshop paper develops new versions of the CORI and KL-divergence resource selection algorithms that better compensate for skewed distributions of database sizes. The three resource selection algorithms are compared on several testbeds with varying distributions of database sizes and relevant documents. The extended version of KL-divergence is about as effective as the new ReDDE algorithm. The extended CORI algorithm is better than the basic CORI algorithm, but is the least effective of the three.

Decision-theoretic resource selection for different data types in MIND by Nottelmann and Fuhr is also a companion to a paper that appeared in the main SIGIR conference. The conference paper reports on a decision-theoretic framework for text resource selection based on characteristics such as relevance, access time, and access costs. The workshop paper extends the approach to other data types, such as person name, year, and image, and exact-match and approximate-match retrieval methods. Two of the three methods presented for estimating the retrieval quality of a digital library can be applied to text and non-text data types. In spite of its generality, so far this approach to federated search of multimedia digital libraries has only been evaluated using text resources due to a lack of large, widely available multimedia resources.

Distributed Web search as a stochastic game by Khoussainov and Kushmerick addresses the problem of maximizing the performance ("profits") of a search service in an environment containing competing search services. Search engines are assumed to compete by deciding which markets to serve with their finite resources, and consumers are assumed to flock to the search engines that best meet their needs. This process can be modelled as a stochastic game in which parties have only partial information, there is a limited range of actions, and actions take time to have effects. Evaluations were done using data derived from 100 days of Web proxy logs from a large ISP; 47 search engines were involved. Experimental results indicate the effectiveness of the general approach, but also

demonstrate artifacts due to assumptions. Future research will be on improving the models, and reducing the simplifying assumptions.

Collection fusion for distributed image retrieval by Berretti, Del Bimbo, and Pala describes a model-based approach to image data fusion (i.e., merging results from different image libraries). During an offline model learning stage training data is acquired by a form of query-based sampling in which queries are submitted to an image library, images are retrieved (with their library-specific scores), and normalized, library-independent scores are computed with a fusion search engine. When sampling is complete, images from each library are clustered into groups, and pairs of library-specific and normalized scores are combined to learn group-specific linear models. During interactive retrieval an image's score is normalized by finding the most similar cluster and using the model parameters associated with that cluster. The method is very fast and worked well in experimental evaluations.

New methods of results merging for distributed information retrieval by Wu, Crestani, and Gibb addresses the problem of merging results exploiting overlaps in different retrieval sets in order to achieve better performance. The new results-merging algorithms proposed take advantage of the use of duplicate documents in two ways: in one case they correlate the scores from different results, in the other they regard duplicates as increasing evidence of being relevant to the given query. An extensive experimentation suggests that these methods are effective.

Recent results on fusion of effective retrieval strategies in the same information retrieval system by Beitzel, Jensen, Chowdhury, Grossman, Goharian, and Frieder takes a new look at metasearch by studying it within a single retrieval system. Metasearch is known to improve retrieval results, but prior research often focused on fusion from different retrieval systems, which conflates effects due to different representations and retrieval models. In this study the representation is held constant. The results are unexpected. The number of documents that appear in multiple retrieval lists ("overlap documents") is considered a good clue to the effectiveness of data fusion; for example, the well-known CombMNZ method exploits this feature. However, it is a poor predictor in this setting, rewarding common "near miss" documents and penalizing "maverick" relevant documents found using only a single method. This paper encourages a more careful examination of representation vs. retrieval model effects in future metasearch research.

The MIND architecture for heterogeneous multimedia federated digital libraries by Nottelmann and Fuhr presents an architecture for distributed information retrieval. It consists of five types of components: graphical user interfaces, data fusion components, a dispatcher, proxies, and digital libraries. Proxies provide "wrapper" functionality for each digital library, providing common schemas and APIs for heterogeneous, multimedia, and possibly uncooperative digital libraries. Proxies also provide local resource selection using a cost-based, probabilistic framework, so retrieval scores are normalized across different media and digital libraries. The architecture provides varying levels of distribution, depending upon user needs. Communication among architecture components is performed using the SOAP protocol. An implementation is available.

Apoidea: a decentralized peer-to-peer architecture for crawling the World Wide Web by Singh, Srivatsa, Liu, and Miller describes a new spider architecture based on dynamic hash tables. Each node is responsible for a portion of the address (URL) space; each domain is covered by a single node, which keeps communication among nodes down to a manageable level. Exact duplicate detection is handled in a similar manner, by converting Web pages to hash values and making each peer responsible for a portion of the hash address space. The distributed approach makes it easy to distribute crawling geographically, possibly reducing communications costs. Initial experiments show very nearly linear scale-up as the number of nodes is increased.

Towards virtual knowledge communities in peer-to-peer networks by Gnasa, Alda, Grigull, and Cremers describes a peer-to-peer architecture consisting of personal digital libraries ("personal search memory" or PeerSy) and an architecture that lets them organize into virtual knowledge communities (VKCs). Virtual knowledge communities occur by clustering nodes based on each node's frequently asked and seldom asked queries and bookmarked documents (considered relevant). New queries are sent to one's personal digital library (PeerSy), one's virtual knowledge community, and Google. The expectation is that documents found within a person's personal digital library and virtual knowledge community will be better matches for an information need, possibly reflecting some combination of past searching and browsing behaviour. The work is at the initial prototype stage.

The personalized, collaborative digital library environment CYCLADES and its collections management by Candela and Straccia describes the CYCLADES system. In this system a digital library is not just an information resource where users submit queries to satisfy information needs, it is also a personalized collaborative working and meeting space. In this space users sharing common interests may organize the information space according to their own, subjective view. They may also build communities, become aware of each other, exchange information and knowledge with other users, and get recommendations based on preference patterns.

Considerations

Workshops on distributed information retrieval were held in conjunction with SIGIR 1996 and 1997 [1]. The response to the 2003 workshop indicates that many of the same issues remain important: for example, data gathering, resource selection, data fusion, and architectures. However, comparison with the earlier workshops also indicates that the topic has matured considerably. Assumptions about small numbers of cooperating, homogeneous resources running the same software are no longer pervasive. Resource selection algorithms are more accurate, more robust, and are beginning to really address multimedia data; fusion algorithms

[1] See `http://www.cs.cmu.edu/~callan/Workshops/nir96/` and `http://www.cs.cmu.edu/~callan/Workshops/nir97/`.

are much less ad hoc and much more effective; peer-to-peer architectures have emerged; and software architectures have become more detailed and realistic.

During the general discussion there was considerable debate about the state of resource selection research. Resource selection has been the driving topic in this research area for the last decade, and there has been steady improvement, but the upper bound remains unknown. Precision-oriented methods dominated past research, with much success, but high recall and high diversity are neglected topics that are particularly important in some domains, for example to better represent the range of information available.

Participants felt that data fusion research needs to continue its transition to stronger theoretical models. The field does not yet understand how differing levels of overlap among resources affect fusion algorithms; the research community is split into "much overlap" (e.g., metasearch) or "little overlap" (e.g., distributed IR), but the real world is more complex. The field also needs to learn to model the interaction between resource selection and data fusion. Improvements in resource selection may have a large effect or none depending on the data fusion algorithm, but today the interaction is unpredictable.

The topic that generated the most discussion was, of course, evaluation. There was broad agreement that there is too much focus on testbeds based on TREC data. Participants felt that it is especially necessary to model the size and relevance distributions of real sites, and that it might be possible to get such information from industry. There was recognition that different tasks and environments will have different characteristics, and that the research community needs to devote more effort to understanding what they are. A major obstacle for many researchers is that distributed IR is still rare in the "real world," so it is difficult to find "real" data, users, and failures. The clearest example of distributed IR in use today is in peer-to-peer networks such as KaZaA.

Relevance-based ranking (RBR) is a convenient and clear metric, but participants felt that the field will need to transition to a utility-based metric, possibly something like Norbert Fuhr's decision theoretic framework, that encompasses a wider range of user criteria. Such a transition will require a much better understanding of user information needs in distributed environments: for example, the importance of relevance vs. communication time, search vs. browsing, and relevance vs. diversity.

One could summarize the discussion of evaluation as a strong worry that researchers are stuck searching under the same old lampposts due to a lack of realistic data and user information needs. Participants expressed support for a TREC track or INEX-style project to focus attention on creating new datasets, task models, and evaluation metrics. The TREC-4 and TREC-5 Database Merging tracks were conducted before a distributed IR research community had developed, and hence they attracted little participation. Today, with active interest in distributed IR and federated search from a variety of research communities, a similar effort would have a much better chance of success.

Acknowledgements

We thank the organizers of SIGIR 2003 for their support of the workshop. We also thank the members of the Program Committee (Donatella Castelli, Jim French, Norbert Fuhr, Luis Gravano, Umberto Straccia) for their efforts on behalf of the workshop. The workshop was sponsored in part by the MIND Project (EU research contract IST-2000-26061), the University of Toronto, and the Knowledge Media Design Institute.

November 2003

Jamie Callan
Fabio Crestani
Mark Sanderson

Organizing Committee
SIGIR 2003 Workshop on
Distributed Information Retrieval